Module 1

I) Data Science Terminology:

- Data Scientist
- Data Analyst
- Business Analyst
- Data Engineer
- Data Governance
- Data Set
- Data Wrangling
- Data Modeling
- Data Mining
- Data Visualization
- Big Data
- Machine Learning

II) What makes a good data scientist:

Personal qualities:

- curiosity
- analytical
- ethical

Essential skills:

- statistics/math
- programming
- communication
- data management

III) Statistics:

Statistics is the science of collecting, organizing, summarizing, and analyzing data to answer questions and/or draw conclusions.

<u>We use statistics:</u>

- to satisfy our curiosity
  - exploring the world around us
  - searching for patterns to lead to discoveries
- to make sure that we can stand on our legs
  - evidence to show that we are right (or wrong)

<u>Statistics rests on two major concepts:</u>

- ***variation***
  - differences or changes in an item
- ***data***
  - observations gathered to draw conclusions
  - context matters

Context matters — always ask:

- ***who*** — describe the individuals who were surveyed
- ***what*** — determine what is being measured
- ***when*** — when was the research conducted?
- ***where*** — where was the research conducted?
- ***why*** — what was the purpose of the survey or experiment?
- ***how*** — describe how the survey or experiment was conducted

IV) Data types:

Data is the information or a set of values collected from surveys, experiments, observations, etc.

In statistics, we classify data into four categories:

- ***nominal*** — labels; no quantitative value; can be grouped
- ***ordinal*** — non-numerical values; can be ranked
- ***interval*** — numerical values; equal distance between; known order and differences
- ***ratio*** — can be compared

V) Statistics types:

- ***descriptive statistics*** — summarizing data
- ***inferential statistics*** — making inferences; determine relationships

A ***population*** is the entire set (of interest).
A ***sample*** is a subset of a population.
Random selection — all items have equal chance to be selected.

VI) Central tendency:

Distribution shows all values in a data set and their frequency.

Central tendency is a value that describes the center or central location of a data set.

There are three ways to describe central tendency:

- ***mean*** is the numerical average of the data set:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} X_i \qquad \text{(for a population)},$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \text{(for a sample)};$$

- ***median*** is the score at 50 percentile, i.e. the number in the middle;
- ***mode*** is the most frequently occurring, the most common number.

VII) Misleading statistics:

- ***Trident sugarless gum***
- ***Colgate toothpaste***

In both cases a list was actually recommended.

VIII) Central tendency preference:

- ***mode*** — nominal data (outliers are fine)
- ***mean*** — interval/ratio data (data should not be excessively skewed)
- ***median*** — ordinal data (skewed data is fine)

IX) Standard deviation:

Standard deviation measures the average distance from the mean.
Standard deviation for the population is computed using the formula

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (x_i - \mu)^2}\,.$$

Standard deviation for a sample is computed using the formula

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N} (x_i - \bar{x})^2}\,.$$

Variance for the population is σ^2. ***Variance for a sample*** is s^2.

X) Standard deviation and variance empirical rule:

The ***68—95—99.7 rule*** states that a random point with normal distribution

- belongs to the interval $(\mu - \sigma, \mu + \sigma)$ with probability around 0.68;
- belongs to the interval $(\mu - 2\sigma, \mu + 2\sigma)$ with probability around 0.95;
- belongs to the interval $(\mu - 3\sigma, \mu + 3\sigma)$ with probability around 0.997.

XI) *Z*-score:

The ***Z-score*** describes the location of a raw value in relations to the mean and standard deviation. It is given by the formula

$$Z_X = \frac{X - \mu}{\sigma}.$$

XII) *z*-distribution and t-distribution:

z-distribution is the standard normal distribution, i.e. the normal distribution with zero mean and unit variance. If $X_1, \ldots, X_n$ are independent identically distributed random variables with normal distribution, then the random variable

$$\frac{\overline{X} - \mu}{\sigma / \sqrt{n}}$$

has z-distribution.

t-distribution is the distribution of the random variable

$$\frac{\overline{X} - \mu}{s / \sqrt{n}},$$

where s is the sample standard deviation.

If the standard deviation is known, then we use the z-distribution. If it is unknown, then we use its estimate s and then the t-distribution. However, when the sample size n goes to infinity, the t-distribution converges to the z-distribution. Therefore, if n is large enough (30 or more), we can use the z-distribution, instead of the t-distribution, even when the standard deviation is unknown. For relatively small n, we should use only the t -distribution.